The innovations in the realm of Artificial Intelligence (AI) and Machine Learning (ML) and their application in the real world come with significant challenges. For instance, a research study on scaling laws for neural language models highlights that the increasing complexity and scale of ML models, coupled with the exponential growth of data, require more powerful hardware and software solutions. This presents an opportunity to develop secure and efficient distributed training algorithms, infrastructure, and frameworks for utilizing large-scale computing resources to train complex models, while maximizing performance and minimizing latency, energy, and computational costs. Many challenges need to be addressed in the next few years, and I aim to pursue a PhD program to address these issues.

My undergraduate degree in Mechanical Engineering provided a strong foundation in hardware, systems, and programming. It covered topics such as C/C++, numerical methods, computer graphics, mechatronics, data structures and algorithms. This foundation and driven by the groundbreaking work of leading researchers in nonlinear dynamic systems, I embarked on an exciting research journey with my undergraduate advisor, Dr. Rabindra Kumar Behera. I worked on the dynamic analysis of beams, formulating a nonlinear geometric model using Euler-Bernoulli beam theory. This involved applying Hamilton's principles and developing a discretized system for numerical analysis and perturbation theory to solve the analytical model. With MATLAB ODE optimization tools including vectorization, symbolic computation, and batch processing to account for resource intensive sections of the model, I successfully simulated and visualized system behavior, comparing my results with existing analytical solutions. This research, which culminated in a co-authored publication, opened my eyes to the profound impact of data analysis and algorithm development in real-world engineering applications. As I worked on this acoustics project and witnessed the power of computational tools in solving complex, resource-intensive models, I discovered that understanding how software optimizations to enhance the performance of hardware systems in solving computational tasks was an area I wanted to explore further.

Motivated by this experience, I pursued a Master's specializing in AI/ML. In my graduate studies, I deepened my expertise in AI/ML through research-based courses on Computer Vision, Natural Language Processing and Deep Learning, focusing on high-performance computing and parallel systems. One of the notable team projects was a research article literature study on Aspect-Based Sentiment Analysis (ABSA) with Supervised Contrastive Pre-Training Learning (SCAPT) to build a language model implicit sentiment classifier. Our goal was to investigate and reproduce the paper's implementation of transformer encoder-decoder with BERT model and SCAPT pre-training architecture. The original model trained 80 epochs with each pre-training consisting of on average 1.5M data samples aggregating in total to few days of training. Due to limited hardware resources, we addressed this challenge and reduced the training to 8 to 10 hours by optimizing the number of epochs, batch size and corpora with larger learning rate and hyperparameter tuning, introducing checkpoints, batch training and utilized cloud GPU clusters to determine the efficient set of parameters, achieving an average test accuracy of 86% on a new dataset which substantiated researchers results.

Another notable project was optimizing a stereoscopic depth perception algorithm using parallel computing techniques for computer vision applications. Initially implemented sequentially, I then refined to parallel algorithm to improve its performance, particularly in handling memory bandwidth limitations, thread synchronization, memory access optimization, edge cases like halo cells and GPU architecture constraints. By leveraging advanced CUDA features, such as coalesced memory access, algorithm optimization, fine-tuning thread block dimensions and CUDA graphs, I achieved significant speedups while improving the accuracy and efficiency of the algorithm. Through careful analysis of performance metrics using CUDA's profiling and timing tools, I optimized the warp size, addressed work scheduling issues and race conditions, and debugged the parallel code. These efforts deepened my understanding of GPU's potential and heterogeneous systems. This project gave me valuable insights into essential parallel processing concepts, such as multi-dimensional

grid-stride techniques, the SIMT model, wrapper-kernel-device methods, lazy vs. eager compilation, identifying bottlenecks and performance-critical sections, GPU architecture, memory hierarchy, and performance analysis. Additionally, I gained hands-on experience with various parallel programming paradigms, including multiprocessing, multithreading, map-stencil-reduction computations, shared and global memory, as well as expertise in building subroutines and kernels with CUDA, C++ and Julia programming. This deep dive into the parallel computing project further sparked my interest in parallel and distributed systems for ML, particularly in the context of hardware accelerators and heterogeneous systems.

My experience goes beyond research and coursework. As a ML Intern at Chubb Insurance, I developed an anomaly detection model using a probabilistic unsupervised learning approach for an enterprise-level software application. I conducted a comprehensive literature review of PyOD paper, benchmarking anomaly detection models, including linear, proximity-based, ensemble, neural net, and graph-based approaches. Based on this, I selected an empirical cumulative distribution-based outlier detection model, which excelled at handling large, diverse datasets, offering scalability and higher efficiency, achieving on average 83% accuracy in anomaly detection over a 0.5M data samples. To account for the challenge of large size of dataset, I utilized Spark data processing tool due to its capability to handle data in-memory for faster computation and additionally the model was deployed on a distributed computing cloud platform in clusters with shared memory access and workload parallelization with real-time monitoring of resource utilization.

My academic journey, starting with my undergraduate in Mechanical Engineering and advancing to a Master's in Data Science and Machine Learning specialization, has fueled my interest in the intersection of hardware and software. I am drawn to research in hardware-software co-design, parallel and distributed systems, and machine learning, particularly in optimizing large language-vision (LLVM) models and high performance computing for various resource efficient such as GPU's and resource-constrained systems such as edge computing. My goal is to explore how hardware acceleration and learning algorithm optimization with better software solutions can improve LLVM models' performance and memory utilization. Additionally, I aim to address challenges in designing efficient hardware accelerators for distributed ML systems, balancing performance, energy efficiency, and computational cost. Moreover, I am interested in exploring robust security and privacy techniques in distributed ML systems. A PhD in Computer Science will enable me to pursue these research interests, contribute to AI's adoption across industries to address societal challenges like environmental sustainability and data privacy, and push the boundaries of ML systems. I aspire to advance ML algorithms and hardware architectures, bridging the gap between hardware, software, and security, while leveraging open-source projects and contributing to original research.

Through collaborative interactions with advisors and peers, which I now see as a critical foundation for research, I have developed a strong foundation in research methodology and a keen eye for identifying and addressing research gaps. My ability to recognize and leverage subtle insights with adapting to evolving research challenges, coupled with persistence in overcoming challenges and embracing diverse perspectives, has led to significant breakthroughs in my research. Having lived in three countries, I also bring a global perspective on technology's potential to transform lives. I believe that my readiness for graduate studies is demonstrated not only by my technical expertise but also by my ability to collaborate across disciplines and persist in overcoming challenges. I have learned the importance of flexibility and perseverance, adapting to evolving research demands, and embracing diverse perspectives to find innovative solutions. In the long-term, I see myself conducting research and leading research teams to pursue my passion for research and mentoring. By contributing to industry and academic research, I aim to address complex challenges and develop impactful solutions in computer systems for ML.